Database Construction

The methods used to construct the database included: locating data and references, defining data types, screening and entry of data, editing and validation of data, placement of data into a geographic and physiographic context, and transfer of information to users of the data. Please read the following text for details.

Collaboration

This database of existing data on chemical contaminant concentrations in sediment for the Gulf of Maine region was compiled with the collaboration and cooperation of many scientists, agencies, and institutions. The participation of the research and regulatory community in defining communal goals, in determining what measurements were important to record, and in assessing how to judge the quality of the rescued data results in products that meet the needs of the Gulf of Maine community. A listing of parameters to include in the database was agreed on and training in data screening and entry was provided to the participants. Principal collaborators, and their assistants or students, were responsible for locating references and entering data within their geographic or topic area. The compiled entries were reviewed as a batch by USGS staff for completeness and quality using iterative validation and screening methods (Manheim and Hathaway, 1991; Manheim et al., 1998). Entries that were identified by the validation process as questionable, data that needed repair, and samples with sparse documentation of quality criteria were reviewed again and appropriate comments were made in the database about these samples. Each collaborator was familiar with the content and structure of the database and could serve as a resource for others in the region on how to utilize searches, graphical displays, and comments to select and use data for specific needs.

Data and references

Data contained in the database originated from many sources (**Table 1**). The USGS. completed searches of existing bibliographies and electronic searches of the American Geological Institute's Geoscience Database (GEOREF), the Aquatic Sciences and Fisheries Abstracts (ASFA), and the National Technical Information Search (NTIS) listings. The ASFA and GEOREF searches identified most of the papers in the peer-reviewed literature that contained significant amounts of data. The NTIS search identified many governmental agency documents that have limited distribution. Such in-house and consultant reports are commonly referred to as "gray literature". Keywords used for the searches included major locations, elements or compounds, and likely general terms. Records held in existing bibliographies, funding agencies, institutions, libraries, and individual contact with scientists and regulators working in marine sciences throughout the Gulf of Maine were used to identify additional documents likely to contain data on contaminants in sediments. Bibliographies reviewed for documents include Regional Association for Research on the Gulf of Maine (RARGOM, 1997), Massachusetts Bays (Massachusetts Institute of Technology (MIT) Sea Grant for Coastal Resources, n.d.), Great Bay (Ward and Pope, 1994), and Bay of Fundy collections (Conservation Council

of New Brunswick, 1993). When an existing compilation of historical data was available (e.g., Metcalf & Eddy, 1984; Cahill and Imbalzano, 1991), the data was transferred electronically and verified with the original data source when possible. Data held in agency databases was also transferred electronically, and associated information about data quality was acquired from published documents and discussions with scientists at these agencies. Agency databases that were utilized include: the NOAA Status and Trends Program (NOAA, 1988), the Massachusetts Water Resources Authority´s Monitoring Program and the US Army Corps of Engineers permit and dredging programs (New England District, Concord, MA, (Buchholtz ten Brink and others, 1992). Documents containing data included in the database were cited in each data table (under "Source of Information or Reference") and full bibliographic references are given in the References. The database contains linked information to aid users in locating original data sources and paper copies are archived at the U.S. Geological Survey in Woods Hole, MA. The compiled bibliographic information also includes related references that did not contain original data on contaminants in sediments.

Table 1. *Location Data and References*

Types of Data Sources		***Data Location Techniques***
Easily accessible sources	***Grey literature sources***	Bibliographic abstract searches
Published journal articles	Technical reports	Monitoring agency queries
Compiled databases	Student thesis	Institutional student records
Monitoring programs	Unpublished project results	Personal contact of specialists
Scientists' own records	Permit applications	Funding agency queries

Measurements of major elements, trace elements, metals, or organic contaminant compounds on whole sediments within the Gulf of Maine were compiled. Those for measurements in sediment fractions, waters, pore waters, or biota were not. The geographic area for sample inclusion is the marine region bounded on the south by Cape Cod, MA., on the east by Georges Bank, on the north by Nova Scotia, and on the west by coastal New England. Some references containing samples in contiguous wetlands, river estuaries, Georges Bank, and the Bay of Fundy were collected; however not all samples from these peripheral areas were entered in this edition of the database nor was the literature scrutinized for data from these areas. The Database of Contaminated Sediments for the Gulf of Maine (Vol. 1) has attempted to comprehensively retrieve analytical data for sediment samples collected from 1950 through 1995; some omissions are inevitable. Data sets for more recent samples (some through 1998) that could be transferred electronically are included; however, newer documents that require hand-entry into the database are not in the current compilation. We maintain a listing of potential data sources and we ask that omissions, mistakes, supplementary information, and new data be brought to our attention.

Ancillary data

In addition to discrete contaminant measurements, the database includes documentation about sample collection, analytical methods, and other information that is required to assess the quality of the reported data. The heterogeneity of the data sources has resulted in a wide range of accuracy and precision for the data that is compiled. Scientific editing of the data (see Data Validation section, below) has identified some clerical or omission problems and permitted many of them to be repaired. Commentary and qualifier information is provided throughout the database to assist users in deciding which data are appropriate for their specific application.

The database targets contaminant measurements on whole sediment (see parameter lists); however, it also records the existence of related chemical data that are not compiled in this document. These data include analysis of size-separate fractions of sediments, special leachate studies, elutriate tests, interstitial waters extracted from the sediments, suspended material or bottom water. Other complementary data that affect the distribution and mobility of contaminants, the toxicity of the sediments, and the capacity of the sediments to sequester contaminants may include bioassays, benthic ecology, contaminant concentration in benthic organisms, biological parameters, habitat classifications, geophysical data and physical oceanographic data.

Database Structure

The Contaminated Sediment Database has a flat-file (spreadsheet) structure, with samples in the vertical dimension and properties in the horizontal dimension. The database is subdivided into six data tables in order to accommodate more than 800 fields without exceeding spreadsheet limitations. Each sample in the database occupies a record (row). Each sample record is linked across the tables by a unique identification number (Sample ID) that is assigned when the data is entered, and by a citation to the original source. This structure is flexible. It allowed unlimited addition of fields as new data types were encountered. It also provided a single structure for data entry, for data processing, and for data output in a format suited for immediate data plotting and evaluation using widely-accessible commercial software. Requirements for special database management skills were minimized. The flat-file structure maximizes flexibility and transportability at the expense of compactness and structured query capabilities. Since software and data manipulation capabilities are changing rapidly, the database in its present structure can be imported into database management software of choice by the user.

Data Dictionary and Database Tables

Data Dictionary

The Data Dictionary defines the parameters that are in each data field included in the six data tables (**Table 2**). These tables contain information about the sample location and collection, measurements in sediments of inorganic chemicals, general organic compounds, polychlorinated biphenyls (PCB) and pesticides, polyaromatic hydrocarbons

(PAHs), and grain size. These linked tables are supplemented by separate glossary and reference tables. The glossary includes abbreviations, methods and devices, and other lists compiled during the construction of the database. The full Data_Dictionary, in vertical format, provides field names for each parameter in three columns, with short field name (10 characters), medium field name (25 characters) and a definition of the field. This choice of format is provided to accommodate restrictions that may be imposed by a variety of software types that are used in the community. The fields within each table, and their full definitions in the Data Dictionary, are organized by subcategory, and are further organized alphabetically within subcategories. ***The Data Dictionary is a working and evolving document that provides detailed definitions of parameter fields, codes and abbreviations. It is suggested that the user print these files and keep them handy while inspecting or extracting data.***

Table 2. *Organization of the Data Tables in the Contaminated Sediments Database*

STATION TABLE	Fields which cite sponsoring and other organizations, field operation data, locations, sampling systems, references and supporting documentation, checklists of the type and numbers of data contained in the database, related information in the references.
INORGANICS TABLE	Information about the sample, analytical methods, and measured values for major elements, trace metals, and other inorganic parameters. All elements have columns for qualifier and detection limit values associated with the concentration field that are not listed here. Some elements also have fields for original values in original units (and original units) if they were not reported.
GENERAL ORGANICS TABLE	Information about the sample, analytical methods, and measured values for the most common measurements of bulk properties of organic contaminants. Data for C, H, or N is in the inorganics table.
PCBs AND PESTICIDES TABLE	Information about the sample, analytical methods, and measured values for major polychlorinated biphenyls and pesticides. Each of the PCBs and pesticides has its own qualifier and, when needed, detection limit column.
PAHs TABLE	Information about the sample, analytical methods, and measured values for major polyaromatic hydrocarbons (PAHs). Each of the PAHs has its own qualifier and, when needed, detection limit column, however, they are not all listed here. Analytical information for data in this table is located within the laboratory information section of the PCBs and Pesticides Table.
TEXTURE TABLE	Information about sediment grain size and lithology.

Information preservation

This compilation aims to preserve the information that is reported in the original references yet make it homogenous enough to compile and manipulate. Most text fields in the database accept unrestricted entry (except for text length) and there are numerous fields throughout the tables for qualifiers and comments about the data and the sample. The Working Dictionary and the Glossary (the alphabetized Working Dictionary) are metadata for the Data Dictionary. They were used to record abbreviations, types of methods or devices used, new parameters, data-entry logs, codes, and similar tables about the descriptive information entered into the database during compilation. Entries were

assigned for a limited number of interpretive and coded fields in order to aid in comparing heterogeneous data. For example, "collection depth" separates "surface samples", which are defined as having more than 80% of their length above 6 cm in depth, from subsurface samples and samples with unknown depth. All available information was used to assign coded fields: geographic location (Area Code), depth in sediment (Depth Code), sampling device (Core or Grab), type of analysis recorded in the Database (Metals & Other Inorganics, Organic Contaminants, Grain Sizes) and availability of related data (Bioassay Data, Other Analysis, Other references). The "row number" field, which is present at the beginning of each table, is used for organization and sorting and can be changed by the user.

Most parameters have multiple fields; e.g., chemical parameters have a field for the concentration value in a specific unit, a field for the detection limit of the analysis, and a field for noting any qualifiers or comments about the measured concentration. The qualifier column records any succinct information that pertains to the specific analysis while more extended commentary can be included in the "comments" field. To protect against inadvertent alteration of the data, multiple fields are also present for parameters that frequently have special formatting, e.g., dates and latitude/longitude. We have converted all measured data to common units and also retained the original raw data formats. Data has been recorded in the Database when it is present in the Source or Reference Document, if it could be located elsewhere, or if it could be surmised without a doubt from information given in the source document. No entries are made (BLANK CELLS) where the data were not reported and could not be found. If a measurement was attempted but could not be quantified, a zero (ZERO) was entered and additional information is usually present in the qualifier fields for the parameter.

The contents of most fields in the Database are suggested by their names, and all fields are fully defined in the Data Dictionary. The following comments focus on selected fields in the tables that are especially important or need explanation.

Station data: *sample identity, location, and documented source*

The most critical fields in the data tables are: (1) the "Unique Sample Identification Number" (Unique ID#); and (2) the "Source of Information or Reference". This Unique ID# was assigned for each sample and is the identifying number through which information for a specific sample is tracked and linked in all the basic data tables. Sample identification numbers from earlier compilations are preserved in the Station Table for reference. The short citation for every sample under "Source of Information or Reference" expedites searches and links the data with the full references, which are provided in the Reference table. Sample collection and analytical schemes may include replicates taken as resamplings at a common location, subsampling of sediments, or analytical replicates from a given sample. Separate Unique ID# values were assigned when data from a common sample are reported in different references or source documents. Replicate numbers were assigned to separate sediment analyses of a common sample. On the other hand, a single Unique ID# was assigned where samples were combined prior to analysis and a compositing scheme was available. If reported, the

identification number given to a sample at collection time or by the original researcher was also recorded in the Database, along with the ship, cruise, device, date, time, depth, and compositing scheme. This information can be useful in locating ancillary information about the sample that may be unpublished, in other documents, or from other phases of a project.

The accuracy of locations and times reported for sample collection varied greatly. Latitude and longitude coordinates are necessary for mapping data; however, their absence does not negate the value of other data reported for a sample. When numerical location data were not available, decimal latitudes and longitudes were estimated from maps or other information (see Data Compilation). Any interpretation of mapped data should consequently utilize the location qualifier fields to understand the limitations of the spatial information. In addition to the citation given in the "Source of Information or Reference", the paper-trail information that was compiled from the documents included sponsoring, contracting, and subcontracting parties, names or locations of projects, other sample names used, related work, the date of data entry, identity of the compiler, and comments about the content or location of the reference. Citations that provided additional details about related studies, methodology, particular analysis, or additional sources of information about a sample occur in commentary fields within all of the data tables.

Analytical data: common features

The data tables (Inorganic, General organics, PCB and pesticides, PAHs, and Texture), follow the Station Table and have a common format: The "Unique ID#" and "Source of Information or Reference" fields are at the beginning of each table of analytical data. Next follows specific laboratory and analytical method information that pertains to all or many of the chemical entities reported in the table for a given source. Both instruments and procedures are noted and quality data for groups of compounds may be consolidated here. Last are the analytical data reported for each sample and each parameter´s qualifier fields. Chemical fields usually have a field for concentration values and specific units, a field for detection limit for the method and component, and a qualifier field that may contain quality or other annotations. Qualifiers include notes on measurements than fell below detection limits, reported detection limits, duplicate measurements, corrected measurements, original reported units, questionable values, editorial or data quality notes, and explanatory comments. Associating quality-control data with analytical values decreases the likelihood that information about data quality will be lost or ignored during data retrieval. Measurements that were made but could not be quantified (values were below limit of detection) were entered as zero. Cells were left blank where no data was available.

The data user is STRONGLY ENCOURAGED to review the contents of the data qualifier fields for every parameter and sample that is extracted from the Contaminated Sediments Database prior to its use so that the validity of that data for a specific purpose is considered carefully.

Inorganic data: *major and trace elements, and other inorganic properties*

There are some parameters listed are in the Data Dictionary that have no entries in the Database; e.g., surface area, resistivity, pH, acid volatile sulfides, and radiochemical and isotopic data. These properties can effect the fate and transport of contaminants in marine sediments but the data not identified in the compiled references. Such supporting analyses may have been measured as part of a project but reported in a different reference that was not available at the time of data entry.

Organic data: *changing methods, bulk organic properties, and organic contaminants*

Improvements in analytical methods for organic contaminants over time have resulted in a decrease of broad-scope measurements like "total PCBs" and an increase in analysis for specific organic compounds. The names of organic compounds, such as are reported in the table of polyaromatic hydrocarbons (PAHs), polychlorinated biphenyls (PCBs) and pesticides, are those cited in original data and are arranged categorically and alphabetically. Microbial contaminants and organotins are also recorded in this table but total and organic carbon is recorded in the inorganic data table. Many organic contaminants are known and reported by more than one name; however, the Chemical Abstract Registry Number (CAS #) is also given for compounds whenever possible. Naming protocols may be confusing: specific organic compounds may be reported as total, sums of certain groups, or with names that differ slightly from those listed here. For example, "Fluorene", "C1-Fluorene", "C2-Fluorene", and "Fluorenes" are different measurements. In this database, results are separated where there is ambiguity about their equivalence. *Data users should carefully consider information recorded in the methodology and qualifier sections, consult original sources if necessary, and use caution when comparing organic contaminant data from differing sources and years.*

Texture data: *sediment grain size and lithology*

Information in the texture table can be used to better understand the geologic context in which contaminants are found and the impact which they might have *in situ*. Sediment grain size (texture) data were originally generated by a variety of methods (Poppe et al, 2000) that can result in non-equivalent units for grain-size measurements. The percentages of sediment in gravel, sand, silt, or clay-size classifications were calculated from sieve-size information according to standard geological boundaries (if data allowed) when the breakdowns were not reported in the source documents. Straightforward conversions between geological grain-size norms and those used for many engineering applications are not possible. *Users should consider information recorded in methodology and qualifier sections for samples prior to use of data, consult original sources if necessary, and use caution when comparing grain size data from differing sources*.

References for the Contaminated Sediments Database

Reference Tables provide full bibliographic citations for: 1) sources of compiled data; 2) other references reviewed for data content; 3) documents and bibliographies pertinent to Contaminants in Gulf of Maine Sediments; and 4) references cited in this publication. The Gulf of Maine Database Bibliography lists documents from which data was compiled. The tabular (Excel) file contains both the full citation and the short citation, which is given in the data tables under "Source of Information or Reference". The List of Additional References Reviewed for Data lists additional references that do not contain samples entered in the database but were reviewed for measurements of contaminants for whole sediments from the Gulf of Maine. These include documents that contained: measurements of related parameters but had no contaminant data; measurements of contaminants in biota, waters, or fractionated sediments; samples outside the study area; synthesized data that was previously reported elsewhere; and new reports. Extensive documentation about sub-areas in the Gulf of Maine is available from a number of libraries in the region. Documents that are referred to in the text of this publication, "Contaminated Sediments Database for the Gulf of Maine", are given in the List of Citations in this Publication. *The paper-trail information that is in the station table, reference tables, and data tables may be useful for selecting and evaluating data and for locating the original sources.*

Once a reference was identified and a copy obtained, it was pre-screened for content. An estimate was made of the number of data points in the reference and the condition of the data. Appropriate data was compiled by the authors, and their assistants and students, according to the Procedure for Document Review and Preparation (**Appendix A**) and the Procedure for Data Entry into Database Tables (**Appendix A**), and other training documents. The primary steps followed were: enter bibliographic data; check for redundancy; evaluate the condition and annotate the data and metadata for entry or repair; locate and enter methodology and other qualifier information; transfer or enter quantitative data; and convert, repair or locate data as needed. Data was checked for errors and internal consistency both when samples were entered from the source document and during validation of the compiled data (see below). Separate spreadsheets were maintained by each data-entry person and the entered data was reviewed and combined at the U.S. Geological Survey. Records were kept of all references inspected for data, those having data entered, the person entering data, all attempts made to locate or repair data or metadata, and pertinent samples not entered in this edition of the Database.

Data validation and quality assurance

Data validation occurs both during the screening and entry of data and when a suite of compiled data are reviewed. The major components of the task are: 1) Inspect the reference for completeness of reporting of the sample location, paper-trail citations, sample field data, analytical methods, and measured values; 2) Identify "missing" critical information that is not reported in the reference; 3) Identify potentially "incorrect" entries

in the database or information reported in the reference; 4) Cross -check other sources of information to verify the status of information or data noted as incomplete, missing, or potentially incorrect; 5) Attempt to locate or repair information, or data, from the reference that is verified as missing, incorrect, incomplete, or questionable; and 6) Record the status of repaired, located, incomplete, missing, questionable, incorrect, and corrected data in the database (in the appropriate qualifier or comment field) for every sample that is affected. Sorting, plotting, and mapping techniques provided a fast and powerful means to identify information gaps and data that was outside the norm (i.e., "outliers") (Manheim, et al., 1998). Scientific judgement was then used in deciding how to resolve data gaps, repair data, and comment on the quality of the compiled data. ***An over-riding principle for the database was that data be recorded as reported in the source document, and all corrections, supporting information, and commentary be clearly noted as such.***

Completeness of reporting and missing critical information

Data was compiled from references that were originally created for a variety of purposes. Consequently, there was a wide range in the amount of detail that accompanied the contaminant data. Latitude and longitude (in some form) was reported for 96% of the samples, as was sampling year; whereas only 83% of the samples had information about the depth of sediment that was sampled. The percent of samples having sampling or analytical methods reported was significantly less. Attempts were made to contact originating laboratories, principal investigators, and identify companion publications in order to locate critical information about the methods and accuracy of the sample collection and analysis. The absence of such data precludes use of the contaminant measurements for many applications since differing methodologies (e.g., acid leach vs. total sediment digestion) can generate data that may not be directly comparable, or for which the accuracy differs significantly (e.g., older vs. recent measurements of organic contaminants). Text entries that were made in the methods fields and the parameter qualifier (or comment) fields document what information was given in the reference, note that found elsewhere, and indicate where seriously comprised data occur.

Identification and verification of questionable data

A batch validation technique (Manheim, et al., 1998) was used to identify data that may have been erroneously recorded or not measured correctly. The compiled data was systematically sorted and plotted to aid in identification of outliers. Histograms, ratio plots, and area maps were used to define "normal" sample distributions from the compiled Gulf of Maine samples and also from the NOAA Status and Trends national dataset. Data falling outside the criteria (**Table 3**) were flagged for further inspection. Reasonable explanations for the data were found in some cases, such as extremely high contaminant concentrations found in proximity to a contaminant source, or values with very low detection limits originating in a specialized research laboratory. Sometimes, no explanation or further reason to suspect the data could be found; but more often, a source of error could be identified. In many cases, such as for typographical or conversion errors, the data could be repaired.

Table 3. *Criteria for Identification of Questionable Data*

Station and sampling information
Do samples plot (lat/lon) in stated location?
Are reported significant digits appropriate to location method?
Do dates appear rounded (e.g., 1st of month) or odd (e.g., in future)?
Do dates and locations agree with others in project?
Are sampling device and sample depth compatible?
Do different samples have same reported values (appropriately or not)?
Are samples reported in multiple sources the same for all parameters?
Are units reported and reasonable?
Analytical information
Is methodology and Quality Assurance information reported?
Is detection limit reported and appropriate for method?
Do different samples (especially sequential ones) have same reported values?
Are reported significant digits appropriate?
Are concentrations similar to those of other samples in spatial proximity?
Are concentrations above the limit of quantification?
Are concentrations higher than the normal population distribution?
Are concentrations lower than the normal population distribution?
Are high concentrations near a potential contaminant source?
Are low concentrations in sandy sediments or water samples?
Are major element concentrations typical of marine sediments?
Are reported concentrations potentially detection limit values?
Do parameter ratios fall within the normal population distribution?
Does one sample have same reported values for different parameters?
Are concentrations an order of magnitude (or 3) higher or lower than reasonable?
Reference quality
Is methodology and QA information reported?
Is laboratory reputable?
Is reference carefully written?
Do many samples meet criteria for questionable data?

Repairing data and documentation of data qualifiers

Qualifiers given in the references, such as detection limits or descriptions of collection and analysis, were recorded in the database. Repaired data included samples which had missing information that was subsequently located, samples reported as measured values that were verified to be detection limit entries, samples with unit or format conversion mistakes, and typographical errors. The repaired values were generally placed in the parameter field and the reported value placed in the qualifier field with an explanation. Data confirmed to be of exceedingly poor quality were also placed in the qualifier field. Editorial comments were entered for samples or analysis that triggered criteria for questionable data that could not be resolved or repaired. Representative qualifier comments are shown in **Table 4**. The presence of these comments does not mean that the data cannot be utilized, rather, it indicates that the user should make individual decisions as to whether the sample was collected, analyzed, and reported with an accuracy that is appropriate for the desired application. We have tried to be comprehensive and thorough in identifying data sources, compiling the data, and validating the heterogeneous data

contained in the database. Some omissions or errors are inevitable, though, so we ask that you bring these to our attention.

Table 4. *Data Dictionary for Qualifiers used in Inorganics Table*

Category of data qualifiers	***Examples of qualifiers***	***Definition***
Detection limit	<DL	Data point was measured and reported as below the detection limit
	LQD	Data reported was below the limit of quantitative detection which is calculated from the instrument detection limit (IDL, mg/L) in a formula where LQD (ug/L)= IDL*1000
	no DL reported, these are low and probably <LQD=ND	The detection limit was not reported in the original document. The reported data values are low and may be around the limit of quantification which would render them non-detectable.
Precision	"estimated value" according to orig. ref.	The original reference listed this sample as an estimate
	[Hg val. doubtful, no. same as Cu value]	The mercury value is in question because it is reported as the same value as copper
	The total is lower than Corg.	When calculating carbon content by difference, the equation for total carbon is the sum of organic and inorganic carbon. This denotes that the total carbon reported value is below the value for the organic fraction which is not possible.
High values	high for this dataset	The reported value is abnormally high as seen in the data validation technique. It appeared as a high outlier.
Low values	low value, given in hard copy, no DL reported	The reported value is abnormally low as seen in the data validation technique. It appeared as a low outlier. The data were verified in the original reference (hard copy) and no detection limit was reported either.
Methodology	Calculated by difference from TotalC-Corg	When calculating carbon content by difference, the equation for total carbon is the sum of organic and inorganic carbon. This denotes that the inorganic carbon was calculated with this equation.
	Reported value below limit of quantification (LOQ). Analytical spike outside 85-115% recovery image.	The reported data were below the limit of quantification. An analytical spike was run and was outside of the acceptable limits.
	Reported values are the mean of replicate analysis	The values reported are an average of the replicate analyses performed on the sample
Sample type	Ammonia as N mg/kg (ppm) dry wt.	Ammonia was measured for this sample as nitrogen in units of parts per million by dry weight
Completeness	no data	no data were reported in this document
	no data in orig. database	no data were reported in the original database

Statistics and Standards	s.d. = 1%	This sample had a reported standard deviation of 1%
	Stds 2.1 ug/g	Standards were run for this sample, this standard was 2.1 ug/g.
Units	orig. units in %?	There is a question on the original units and if they are in percent
	original DB also questions this #	the comments in the original database question the value reported
	original units in question	the original units reported for this sample are a concern
"Symbols"	[ ]	database editor comments
	" "	quotes from original document

The data user is STRONGLY ENCOURAGED to review the contents of the data qualifier fields for every parameter and sample that is extracted from the Contaminated Sediments Database prior to its use so that the validity of data for a specific purpose is considered carefully.

Database access and data utilization techniques

This web site provides the description of the Gulf of Maine database project, descriptive plots and maps of compiled data, and access to viewing and down-loading the data tables. All of the data compilation was accomplished with spreadsheet software (usually Excel or Quattro Pro) on both PC and Macintosh platforms. Commercially available database software, such as PARADOX, 4th Dimension, FoxPro, and ACCESS were tested or used at various times to determine if they provided significant advantages for data manipulation and to insure that the data was compatible with a variety of database structures. The plots and maps used for data validation were also generated by an assortment of programs and platforms: Kaleidagraph, Deltagraph, and Sigmaplot; MAPINFO, ARCINFO/VIEW, Grapher, and others. Bibliographic information was also compiled in and converted between various formats. All of the data access and manipulation tasks can be accomplished with minimal investment of software or hardware. The site can be viewed with most common browsers, and is constructed with compatibility for Netscape Communicator Version 4.5 and Internet Explorer Version 4.0. The data dictionary and data tables, which are provided in Microsoft Excel 4.0, can be imported to most word processors, spreadsheet, database, and data manipulation programs. Tables can be viewed, downloaded and manipulated on any computer platform that has appropriate software installed and sufficient memory to open the data tables.

These compiled data are intended to be a resource for researchers and managers in the Gulf of Maine. Potential applications are numerous. They include mapping surficial sediment concentrations to identify potentially toxic areas, assessing the thoroughness of data reporting in regional literature, identifying areas that have a paucity of measurements, determining the scale of necessary monitoring, quantifying changes in environmental conditions over time, locating specific historical samples, selecting

indicator parameters, and others. Selective sorting, plotting, or mapping the data that is compiled in the data tables provides a means to accomplish this.

References

Buchholtz ten Brink, M.R., Manheim, F.T., and Hathaway, J.C., 1992, Results of Boston Harbor Contaminated Sediment Database Development. Report on Cooperative database development with U.S. Army Corps of Engineers, New England District, 11 chapters and electronic media.

Cahill, Jeanne and Karen Imblazano, 1991, An inventory of organic and metal contamination in Massachusetts Bay, Cape Cod Bay, and Boston Harbor Sediments and assessment of regional sediment quality, National Network of Environmental Management Studies Interns, Environmental Protection Agency, Region I, Water Management Division, Massachusetts Bays Program, Final Report.

Manheim, F.T. and Hathaway, J.C., 1991, Polluted sediments in Boston Harbor, Massachusetts Bay: Progress report on the Boston Harbor Data Management File. U.S. Geological Survey Open-File Report 91-331, 27 p.

Manheim, F.T., Buchholtz ten Brink, M.R., E.L. Mecray, E.L., 1998, Recovery and validation of historical sediment quality data approach from coastal and estuarine areas: an integrated approach. Journal of Geochemical Exploration, v. 64, p. 377-393.

Metcalf & Eddy, 1984, Boston Harbor Data Management File, Electronic data file in SAS format, with supplementary hard copy bibliography (prepared for the U.S. Environmental Protection Agency, Region I; coding incomplete; Available from U.S. Environmental Protection Agency, Harbor Studies, JFK Building, Boston MA, 02128.

MWRA (Massachusetts Water Resources Authority), 1999. Harbor and Bay Monitoring program. Online at http://www.mwra.state.ma.us/harbor/html/bhrecov.htm

NOAA (National Oceanic and Atmospheric Administration), 1988, A Summary of Selected Data on Chemical Contaminants in Sediments Collected During 1984, 1985, 1986, and 1987: National Status and Trends Program for Marine Environmental Quality, National Oceanic and Atmospheric Administration (NOAA), National Ocean Service NOS OMA 44.

NOAA (National Oceanic and Atmospheric Administration), 1999. Status and Trends Program, Online at http://seaserver.nos.noaa.gov/projects/nsandt/nsandt.html

Poppe, L.J. and C.F. Polloni (eds.), 2000, USGS East Coast Sediment Analysis:

Procedures, Database and Georeferenced Displays, U.S. Geological Survey Open-File Report 00-358 Online at http://pubs.usgs.gov/of/of00-358/

US Army Corps of Engineers, 1999. New England District Environment Program, Online at http://www.nae.usace.army.mil/environm/environm.htm

APPENDIX A: Training Documents

I. Procedure for Document Review and Preparation
II. Procedure for Data Entry into Database Tables

I. Training Document: Procedure for Document Review and Preparation

Checklist for Document Review and Preparation

1. Prescreen the document.

a. Review the document and earmark the data and text that are to be entered and/or copied for reference. (Highlight or annotate on a photocopy to preserve the original document).
b. Evaluate the general quantity and format of the data in the document. Determine where data tables can be scanned or acquired electronically and where individual keyboard entry is needed.
c. Note the completeness of location data, methodology, project status and any special observations or instructions.
d. Assign a priority level for data entry.

2. Check for redundancy (use successive sorting on key fields).

a. Check whether the samples, measurements, or meta-data in the report are already entered from a differing reference.
b. Check for differing versions or dates of the reference.
c. Determine which of multiple documents, samples, or analysis is to be cited and annotate as needed.

3. Evaluate the condition of the data and metadata.

a. Determine the status of the location data.
b. Determine whether data are in standard units.
c. Identify any data requiring new fields or qualifiers.
d. Note any major gaps in the data, the paper-trail, or the meta-data.
e. Mark or record any items that should be included in "Comments" fields.
f. Record questions or items needing further thought, decisions, or investigation.

4. Enter references in the Project Bibliography.

Include entry of keywords such as the parameters studied, agency, area and lab references, number of records or samples analyzed, and pertinent studies in original report which are not included in database As data are entered, a bibliographic listing of data sources and references must be maintained. Start an Endnote bibliography with all the references you enter. Papers and reports should be entered with all the authors' names, year, title of the report, journal name, or agency name, volume, report name, and any other information pertinent to this reference.

Example:
Moffett, A.M., Poppe, L.J., and Lewis, R.S., 1994. Trace Metal Concentrations in Sediments from Long Island Sound. U.S. Geological Survey Open-File Report 94-620

5. Enter data following the "Procedure for Data Entry into Database Tables"

6. Rescue metadata

a. If no latitude and longitude are listed in the document, proceed to recover the location by digitizing the data from a chart, inquiring from persons involved in the study, or searching in documents about ancillary analysis from the same location. Several techniques are available to recover locations from maps:

- digitize the map and use software (e.g. GIS, or other plotting programs) to assign locations
- scan and "stretch" maps to fit calibrated location maps
- scale latitudes and longitudes (from a copy enlargement) using proportional dividers and standard maps
- prepare a transparent grid overlay using a Gerber scale and/or 10-point dividers
- match features of NS and EW scales to a standard map using copy reductions or enlargements

b. If paper-trail or citation is incomplete, proceed to recover the original document, inquiring from persons involved in the study, e.g., primary authors or agencies. Several places can be checked to attempt to compete a paper-trail or citation:

- scientific libraries and/or journals
- original document source such as author or agency
- online CD-ROM databases
- other persons involved in the study or others who have worked with that particular author or agency

II. Training Document: Procedure for Data Entry into Database Tables

Have the following items handy before beginning data entry:

1. Photocopy (if possible) of the document or file to be entered
2. Highlighter and a pencil
3. Data dictionary
4. Data entry header files (6): Station, Inorganics, General Organics, PCBs and Pesticides, PAHs, and Texture.
5. Working dictionary file or glossary file

There are basically two types of documents that are entered in the database: **(A.) papers and reports** (which includes reports, journal articles, theses papers, unpublished data, etc.) and **(B.)** Army Corps of Engineer **permit files**. The main key to data entry is DETAIL. Enter as much about the data as possible; there are plenty of comment and qualifier fields to place information.

If something seems questionable, or you don't know where a certain piece of data goes, put your question in a comments field (relevant to that data) until you can ask someone for guidance. Also mark the question on the document either by writing it beside the item in question or placing it on a Post-It-Note and sticking it on that page in the document. These flags in the comments fields and on the paper copy let the reviewer know that the item needs further investigation.

(A.) Entry of Papers and Reports

1. Photocopy the document so that you can write on your copy.
2. Photocopy bibliographic data
(or verify that another copy is on file.) Assign a short reference name and enter in both the station table, project bibliography and status table and all subsequent tables.
3. Use the Working Dictionary and Glossary.
Any abbreviations or acronyms that the data entry person encounters or decides to utilize (to speed up the data entry process) should be listed in a glossary. You may use the existing files or create an EXCEL table to serve as your list or Glossary for maintaining these items. The Glossary can be sorted in alphabetical order to look up abbreviations more quickly. The Working Dictionary has categories already set up that different glossary items will probably fall under, such as Agency Names, Sampling Devices, Analytical Methods, etc. Usage of abbreviations vs. complete terms can be made consistent throughout the database at a later stage as long as records are kept and there are no ambiguities.
Example:
The U.S. Army Corps of Engineers is the agency identified as "USACOE." The acronym USACOE can be used in the station table and recorded in the glossary and working dictionary by entering "USACOE" in a column labeled "Abbreviation" and "U.S. Army Corps of Engineers" in a column labeled "Definition."

4. Identify the data
Skim through and find the following items:

What types of data are in the report?
What is the total number of samples, regardless of what is analyzed for each sample?
(This is very important because sometimes there are more samples shown for one type of data than for another, and all samples must be entered whether they contain data or not) As you skim through the document mark off or highlight the items that are to be entered into the data files. Reply on pre-screening commentary when available. You will check and date them when entry is completed.

5. Enter data: station table.
(Name your file with "stat" somewhere in the file name, i.e. STAT1999.xls)

a. Sample Identification

Identify samples in the *Local ID* column by using your initials and a number. For example, if your name were Jamey Reid, you would enter JR1 for the first sample you enter. On the document, write down what your local ID number is for that sample. For example, if sample M-4 is the first sample you enter, and you call it JR1, write JR1 down next to it in the report.

Identification numbers are very important in order to retain original information in the report and they should remain consistent throughout **ALL** data tables. They´re also important when a reviewer attempts to verify the data that has been entered by comparing it to the original document. It provides the ability to point to a sample in the database and find it easily in the report you entered it from. Two important identification numbers are: *Sample ID or Original No.* and *Orig. Station.* These are numbers which can usually be found in the tables within the report or permit file along with the data. The Unique Sample Identifier is an identification number which is assigned by USGS after the completely entered dataset has been reviewed and verified. The format of this Unique Number is US0####, where # are numbers.

(The *Preceding Database* and *Preceding File Name* are used only if you are transferring data that is already in electronic format. Some data may originate from an electronic file taken from an already existing data compilation or database. The name of the database would be entered into *Preceding Database*, e.g., USGS Contaminated Sediments Database and the actual name of the file that was used to copy the data from would be entered into *Preceding File Name*, e.g., data.xls.)

b. Sources of References

The next important field is *the Source of Reference*, which should be the author´s last name first, initials and year. Ex: Moffet, A.M., et al, 1994. (Use ‘et al’ when there are more than two authors.) For two authors, you can use both names. Ex: Esser and Turekian, 1993. For *Project Name*, I recommend using the title of the document because it creates a further link between the database and the hard copy of the document. *Quad*

Name is more relevant to permit files.

c. Location Information

Most reports don't give *State Plane N* and *State Plane E*, but if they do, enter the information in these fields. Look for a table of latitudes and longitudes. (Remember that longitude degrees in the Western hemisphere are entered with a negative sign in front.) If there is no table given, proceed to enter the data. Find latitudes and longitudes from the map when time permits.

General Location Name and *Specific Location Name:* Example: "Long Island Sound" is an example of a general location name, but more specifically the location could be "New Haven Harbor" or "Guilford Marina," etc. Other examples are: a data set is located in the "Connecticut River" and the document says "in the vicinity of the Coast Guard Academy" Therefore "Connecticut River" would be the general location name and "vicinity of the Coast Guard Academy" is the specific location name.

The next part is entry of the *sampling date(s)*, or the date on which the data in the document was collected. This data is usually found in the text of the document or sometimes at the top of a data table. Many documents will only list a Sampling Day/Month/Year 1. ('1' refers to the first day of sampling and '2' refers to the last day of sampling.)

d. Sample Collection Information

Gen. Comment re sample: when more explanation of each sample is given in the text, it should be entered here
Cruise Id - this is usually found within the text of the document either in the introduction or under "Methods." An example is "RV Asterias.
Core or Grab #
Sampling Device: This is usually found under the "Methods" section of a report. It will probably say something like "Samples were collected using…."
Sample Type: should always be sediment, do not enter elutriate (water), or sludge data
Depth in core or sediment: would be found on data table or in text under methods
Core Length: sometimes given in a data table; may also be in methods
Interval Number: not usually given
Depth interval, top of core: if a samples lists 0-6cm, then the "top" value is 0
Depth interval bottom of core: if a samples lists 0-6cm, then the "bottom" value is 6
Orig. depth in sediment: original values are entered here. If the data was originally measure in feet, entered those values here.
Original depth units: the original units of this value such as meters, feet, etc.
Sediment depth code: depth code is either "depth" or "surface;" surface being the first 0-6 centimeters of a core, depth being deeper than 6 cm in a sediment core
Sediment depth comments: notes on the individual sample

e. Information about type of data and data entry

A Yes or No (entered as "Y" or "N") answer should be entered in the following fields: *Metals and other inorganics?, Organic Contams. analyzed?, Grain Size analyzed?* and *Bioassay data available*?
Comments- Bioassay: enter what types of bioassay are in the report here, for example, if there's data on winter flounder and tube worms, etc. enter that here.
Bio reference: if the bioassay data are not in this reference, enter where this data can be found
Other types Analy. In Ref: what other types of data are in this reference? Such as elutriate, sludge, etc.
Data Entry day/month/year/formatted: Enter the whole date into one cell and accept the default format because it may change as files are transferred between different systems, hence the need for separate day, month, year columns. (*Example:* December 29, 1999 may appear as 29-Dec-99)
Initials of Data Enterer: Enter your initials so you can be identified with the data you entered if any questions arise later.

(B.) Entry of Permit Files:

Source or Reference Name for permit files should be entered in the following way:

ACE_NED permit file #regulatory file no., Project Name

Example:
ACE_NED permit file #1995-00138, Groton Long Point Association

Permit files can be very complicated and to enter agency information, you need to sort through to see who did what. For example, Portsmouth Yacht Club may be sponsoring the work, but they hire Braman Engineering to do the work, and Braman Engineering may in turn hire someone else to help do the work.

Agency 1 Sponsoring (agency publishing the work)
Agency 2 Contracted (agency/researcher doing the sampling)
Agency 3 Subcontracted (agency/researcher doing sampling, not analytical labs)
Agency 4 Other (additional agencies responsible for work)
Project Name: included in the name of the file
Quad Name: this is a four-letter abbreviation for the area on the map, and can either be found on a data sheet or on the map itself.
Regulatory File Number: the number on the application, see example above
Est. vol. of material: how much material will be dredged up; measured in cubic yards; is usually found on the permit application
Disposal area code: a four letter code that indicates where the material will be disposed of; can either be found on the data sheet or on the permit application

All other procedures related to papers and reports are must be done with permit files also.

6. Enter data: tables of analytical data

The next sequence of data entry should be **2) inorganics, 3) general organics, 4) specific organics (parts 1 and 2), and 5) texture**. Name the files accordingly

a. Sample identification information

Be sure to use the same local ID number in all of the files and use the same original numbers throughout. For example: if the samples in the report are A, B, C, D, and E, then those sample names should appear in every file either under *Laboratory's sample ID number* or *Laboratory's Job Number* (even if it's not the lab ID) These numbers can usually be found on the lab sheets themselves if they are included in the document or in the text of the document. Also use the same *Source or Reference Name* for all files. If you find that a report includes only inorganics and texture, or only organics, or some other variation, you must enter the data in **ALL** tables (as a space holder) and put in a comment such as "no data of this type for this sample"

Example: "no organics data for this sample."

b. Laboratory and methods information

All of the data tables have columns in which to enter the following information:

Testing Lab: This is the actual name of the testing lab found on the lab sheets themselves or in the text of the document.
Analytical technique: Enter the technique used to analyze this sample as written on the lab sheets or in the text of the document
Analytical comments: If there are any comments or further description pertaining to how this sample was analyzed, they can be entered here
Replicate number and number of replicates:
Test day/month/year: This information is usually found on the lab sheets themselves or in the text of the document.
Test date formatted: Enter the whole date into one cell and accept the default format because it may change as files are transferred between different systems, hence the need for separate day, month, year columns. (*Example:* December 29, 1999 may appear as 29-Dec-99)

c. Measured data values

Measured data values should always be entered in the concentration field for each parameter.

Parameter concentration	Parameter Qualifier	Parameter Detection Limit
Concentration goes here		

Example:

Arsenic (As) mg/g	As qualifier	As detection limit
100		

The values should always be entered in the units that are given in the tables. Most inorganic parameters are measured in mg/g (micrograms per gram) while most organics parameters are measured in ng/g (nanograms per gram). If the units of the data in the document do not match the units in the tables, then the data should be converted. A comment should be made in the qualifier column stating that the data was converted and what units it was originally reported in.

Conversions are as follows:

	These units:	are equal to:	and also equal to:
1.)	µg/g =	ppm =	mg/kg
2.)	ng/g =	ppb =	µg/kg
3.)	1 µg/g (or ppm or mg/kg) =	1000 ng/g (or ppb or µg/kg)	
4.)	1 in. =	2.54 cm	
5.)	10^6 µg =	1 g	
6.)	10^4 µg =	100 g =	

d. Qualifiers, comments and detection limit values

Any comments in the document or data tables pertaining to a particular measurement should be entered in the qualifier column for each parameter. Detection limit values should always be entered in the detection limit column for each parameter. (Detection limit values will be listed in the document.)

The qualifier column should also include any notes or comments found in the text or data tables (within the document) about the sample. *Only numeric values should be entered in the concentration and detection limit columns*. Abbreviations that indicate that a sample is below the detection limit are: "BDL," "ND," "below MDL," etc. If a document states that a sample was a non-detect or below the detection limit yet no detection limit values reported, enter a 0 for the concentration and put an appropriate comment in the qualifier field such as "reported as ND, detection limit values not reported." When no measurement is attempted, the concentration field remains empty.

Samples that are below the detection limit are to have this information entered in 3 fields:

Parameter concentration	Parameter Qualifier	Parameter Detection Limit
0	<	Detection limit value goes here

Examples:

Arsenic (As) m g/g	As qualifier	As detection limit
0	<	.001

Arsenic (As) m g/g	As qualifier	As detection limit
0	reported as ND, detection limit values not reported	